Expectations of Fairness and Trust Co-Evolve in Environments of Partial Information

Rauwolf, Paul; Bryson, Joanna

Dynamic Games and Applications

DOI:
10.1007/s13235-017-0230-x

Published: 01/12/2018

Publisher's PDF, also known as Version of record

Dyn Games Appl
DOI 10.1007/s13235-017-0230-x

Expectations of Fairness and Trust Co-Evolve in Environments of Partial Information

Paul Rauwolf[1,2] · **Joanna J. Bryson**[2]

Abstract When playing one-shot economic games, individuals often blindly trust others, accepting partnerships without any information regarding the trustworthiness of their partner. Consequently, they risk deleterious pacts. Oddly, when individuals do have information about another, they reject partnerships that are not fair, despite the fact that such offers are profitable—individuals costly punish. Why would one reject profitable partnerships on the one hand, but risk unknown offers on the other? Significant research has gone into explaining the contexts where blind trust or costly punishment provides an evolutionary advantage; however, both behaviours are rarely considered in tandem. Here we demonstrate that both behaviours can simultaneously be revenue maximizing. Further, given the plausible condition of partially obscured information and partner choice, trust mediates the generation of costly punishment. This result is important because it demonstrates that the evolutionary viability of trust, fairness, and costly punishment may be linked. The adaptive nature of fairness expectations can best be explained in concert with trust.

Keywords Trust · Partial information · Costly punishment · Fairness · Trust game · Negative pseudo-reciprocity

1 Introduction

The trust game offers an individual the opportunity to invest with another in hopes the other can be trusted to return an increased investment. When offered the opportunity to invest,

Electronic supplementary material The online version of this article (doi:10.1007/s13235-017-0230-x) contains supplementary material, which is available to authorized users.

✉ Paul Rauwolf
p.rauwolf@bangor.ac.uk

1 School of Psychology, Bangor University, Bangor, UK

2 Artificial Models of Natural Intelligence, Department of Computer Science, University of Bath, Bath, UK

Published online: 25 September 2017

Table 1 Trust game

	Defect	Trust
Investor	1	rb
Trustee	0	$(1-r)b$

An investor begins with one unit of fitness. That investor then chooses to trust and invest in a trustee, or to defect. If the investor defects, then they keep their payoff of 1, and the trustee receives 0. If the investor trusts, then the investment is multiplied by some factor b, and the trustee returns some fraction r to the investor. The investor will earn a profit if the trustee returns $r \geq 1/b$

human players tend to make two seemingly poor decisions. First, when individuals possess no information about their partner's trustworthiness, they trust and invest despite the fact that the subgame perfect strategy is for their partner to defect [11,44]. Second, when the trust game is altered such that players know that the amount to be returned will be profitable, they reject the offer unless the trade is fair—they express costly punishment against a potential benefactor [47,49].

Why would it be beneficial to reject profitable offers on the one hand, but accept unknown (and thus potentially deleterious) offers on the other? From a game theoretical perspective, the opposite behaviour is optimal. In a one-shot game, it is better to never blindly accept potentially deleterious deals and to always accept profitable offers. Have blind trust and costly punishment not been vetted by the evolutionary process? Over the last decade, theoretical research has shown that, when taken separately, blind trust [47,48,53,78] or costly punishment [8,26,43,65] can in some contexts provide an evolutionary advantage. Here for the first time, we offer a parsimonious simultaneous account for both blind trust and rejection of unfair offers.

A strategy comprised of both blind trust and costly punishment can in fact be revenue maximizing and thus adaptive. We analyse the one-shot trust game in a context where participants have both partner choice and partial information about partners' trustworthiness. We find that trust and costly punishment can both be adaptive, and that trust mediates the evolution of costly punishment. Trust and costly punishment are revenue maximizing strategies when taken in the context of each other and in environments where information is partially obfuscated. Within evolutionary game theory, the analysis of co-evolving traits has a long tradition [56]. Here we extend and exploit that literature to demonstrate that blind trust and fairness norms are both adaptive when considered in tandem.

2 Motivation

In the trust game (TG), an individual (the investor) is given one unit of money. They can either keep it and walk away, or invest it with another individual (the trustee). If invested, the money is multiplied by some factor, b, and the trustee can choose to return some portion to the investor. The amount returned is determined by the trustee's return rate r; any rate greater than $1/b$ garners the investor a profit (see Table 1). In a one-shot game where the investor does not possess any information about the trustee, defection is the subgame perfect equilibrium. Because the trustee never benefits from returning anything to the investor, the investor should not trust. But humans do tend to trust, and to invest, despite its suboptimality [11]. Attempts to diagnose this propensity for blind trust have spanned both empirical [1,13,19,23,27,29,44,45,57] and theoretical [16,33,50] disciplines.

Recently, the theoretical literature has investigated whether experimentally observed levels of trust can be explained by positing investors with some knowledge of the trustee's return rate. The intuition is that if humans evolved in small groups where reputations are difficult to hide, perhaps the levels of trust observed in experiments would be adaptive. Manapat et al. [47] have shown that selective pressure for trust can arise when the chance of knowing a trustee's return rate exceeds some threshold. This holds even when information is delayed and inconsistent [48], or if information is costly [53]. Further, population structure can increase the contexts where trust is adaptive [78].

Introducing knowledge of individual trustworthiness into the trust game opens up a new way to consider human play. We may now think of the TG as it relates to the ultimatum game (UG) and recognize failures to trust as a form of costly punishment. In the UG, a proposer offers some fraction of a windfall to a responder. The responder decides whether to accept the offered partition or to reject it—in which case neither side keeps any share of the investment. The subgame perfect equilibrium is for the responder to accept any offer; but most UG responders will reject unfair offers [49], though what is deemed 'fair' varies somewhat across cultures [38].

In the TG, if an investor knows the expected return rate r for a trustee, their decision to invest becomes relatable to the responder's decision in the UG [78]; the investor must decide whether to invest with another when they know the eventual return rate. Experimental research has shown that when investors in the trust game are given information about the trustee's return rate, investors often reject offers that would be profitable ($r > 1/b$, [47]). Similar to the UG, rejecting an offer when it provides a net gain for the investor can be interpreted as a form of costly punishment. The investor is willing to pay a cost (i.e. lose potential profit) to reject an offer that is deemed unfair.

Costly punishment is considered key to understanding the level of cooperation expressed in contemporary human society [32,39,60]. Nevertheless, both the ultimate and proximate causes of costly punishment remain an open question [37,66,73]. Some researchers have suggested that costly punishment is explained by a human predisposition towards fairness—that individuals are willing to pay to maintain a norm of fairness [30,32]. If this were the case, we might expect costly punishers should act fairly in other contexts, but evidence suggests costly punishment is not always correlated with prosocial behaviour [15,82]. Experimental data suggests that punishment is expressed for multiple reasons, including both revenge and a predisposition for fairness [14]. Even where it benefits society, punishment is often motivated by anger [42] and is known to be linked to social dominance [58,59]. Evolutionary game theory has shown that costly punishment can be adaptive if selective pressures focus on relative rather than absolute payoffs [8,43], if punishment occurs probabilistically [46], if payoffs for cooperation are nonlinearly related to the number of cooperative partners [62], if punishing generates a reputation which increases cooperation in the future [41,71], or if selective pressure is weak [65].

Adding the perspective of costly punishment to the trust game may only seem to confuse matters further. Why would it be beneficial to reject advantageous offers on the one hand (costly punish), but accept offers of unknown—and thus potentially deleterious—quality on the other (blindly trust)? If humans are predisposed to fairness, then why is costly punishment not correlated with trusting others [82]? If humans are attempting to outcompete each other through relative payoffs [43], then why act trustingly at all [11]?

Partner choice and its resulting market dynamics may solve part of this quandary [21,28,55]. When people can form reputations and select partners, higher levels of cooperation are witnessed in the laboratory [76]. Recent research indicates that expectations of fairness can arise from partner choice [3,24,26,75]. The intuition is that competition to be selected forces

a rise in prosocial behaviour [6,35,55,68]. However, these results still do not explain why an individual would reject all unfair but profitable offers in a one-shot game.

One reason why players may reject all unfair but profitable offers during economic games is because they are using a natural heuristic which assumes they can always seek other, unknown partnerships. Historically, costly punishment has been defined as the propensity to incur a loss in order to punish those who behave unfairly. In the one-shot trust and ultimatum game, this is witnessed in the tendency for players to reject all profitable, but unfair offers, thus garnering less profit for themselves. However, when multiple available partners and some partner knowledge are included in the trust game, behaviours that appear like costly punishment may not be costly to the punisher. If rejecting all known profitable-but-unfair offers and blindly trusting an unknown offer increases one's payoff in the same round, then such behaviour might not be termed 'costly punishment'—the maker of the poor offer is punished, but if the blind offer is expected to be higher then the punisher does not expect to pay a cost. This implies that at least in some conditions, sanctioning behaviour often perceived as costly punishment may be part of a system of negative pseudo-reciprocity [18]. Individuals who are free to seek alternatives may leave profitable, but unfair partnerships in order to seek improved payouts. It is well known that in such environments negative pseudo-reciprocity can explain the rejection of current partnerships [26,52].

What has not been shown is how such behaviour interacts with trustworthiness. Here we present a model which, to our knowledge, is the first to describe how blind trust and risky negative pseudo-reciprocity—which will sometimes manifest as costly punishment—can co-evolve. We find that a combination of both blind trust and a willingness to sanction others creates a higher payoff than either one of those behaviours in isolation. The literature describing the evolutionary functions of trust and costly punishment often presumes that the suboptimal results of one-shot experimental games are a consequence of the fact that participants use heuristics which are adapted to more natural environments. Our work presents a similar but more parsimonious explanation for why both trust and rejecting all known but unfair partnerships may be functional. Punishment which often manifests as costly can be explained in tandem with trust, if we assume individuals typically operate in environments of partial information and partner choice.

By creating competition between (i) trustees whose individual return rates are known and (ii) trustees with unknown individual rates, we find that blind trust and rejecting all known profitable offers can both be adaptive. An individual with high trust and a willingness to sanction earns more profit than those who share only one or neither of those behaviours. Further, we show that rejecting known profitable offers cannot evolve in this context without trust, and that such behaviour is only adaptive in environments of partial information. This form of punishment can be advantageous if and only if it includes the implicit threat to trust unknown offers.

3 The Evolution of Trust

3.1 Model

Manapat et al. [47] argue that the trust demonstrated in the laboratory may be a consequence of humans evolving in small groups with partial information. Here we replicate their results using simulated evolution. We show that trust is adaptive given both partner choice and occasional knowledge of trustees' return rates.

In this model, a population of investors and trustees play multiple one-shot trust games. The population consists of $N_i = 500$ investors and $N_t = 500$ trustees. A trustee is genetically encoded with $r \in [0 \ldots 1]$, its *return rate*. This is the fraction of the investment the trustee will return to the investor. Each investor possesses a *trust* attribute $t \in [0 \ldots 1]$, which represents the chance of the investor trusting a trustee when they possess no information regarding the trustee's return rate. For all simulations shown here, we set $b = 3$. With probability t, an investor trusts a trustee and the trustee is given $b = 3$. The trustee then returns rb to the investor and retains $(1 - r)b$ for itself (see Table 1). With probability $(1 - t)$, the investor does not trust and retains the one unit of fitness, leaving the trustee with nothing.

In a one-shot game where the investor has no information regarding the trustee's rate of return, the subgame perfect equilibrium is not to trust. This changes when an investor (i) may select from one of multiple trustees and (ii) might have some knowledge of a trustee's return rate.

In this model, each round an investor is presented with k randomly selected trustees from the population of N_t trustees. q is the probability of knowing the return rate, r, for a given trustee. The investor may then invest with one of the k trustees. For example, if $k = 4$ and $q = 0.5$, then four random trustees are selected for a given investor. Since $q = 0.5$, there is a 50% chance the investor will know a given trustee's return rate. As such, on average the investor will know the return rates of two trustees, whilst the return rates for the other two will remain unknown. The investor can then select one of the four partners.

How should an investor choose among many potential partners when they have information about some trustees and no information about others? Here we slightly alter the decision rule presented in Manapat et al. [47]. Keeping in mind that the investor only makes money if the trustee's return rate is greater than $1/b$ (in this case $1/3$), we presume that an investor (i) selects the highest known return rate as long as $r > 1/b$. (ii) If no such return rate exists and the investor is trusting, they invest in an unknown return rate, otherwise (iii) they do not invest and keep the 1 unit of fitness.

This can be written formally. In a particular game, the investor will know the return rate of j out of the k trustees, based on the value of q. Consequently, an investor will select trustee i with return rate r_i with probability:

$$p(r_i) = \begin{cases} 1 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x = r_i \geq 1/b \\ 0 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x \neq r_i \\ \frac{t}{k-j} & i > j;\ \max\limits_{1 \leq x \leq j} r_x < 1/b \end{cases} \tag{1}$$

If the return rate, r_i, is the highest known rate, and the return rate is greater than $1/b$, then the investor will select it. If r_i is known, but there are other larger, known return rates, it will not be selected. Finally, if none of the known return rates are greater than $1/b$, then the investor will randomly select an unknown return rate with probability t. Thus, as an investor's trust increases, they are more likely to take a risk and invest in an unknown trustee.

This decision rule is employed for a couple of reasons. First, the investor is profit maximizing when information is known. If the return rates for all trustees are known ($q = 1$), the investor selects the highest return rate, presuming the highest offer is profitable ($r > 1/b$). Second, the decision rule tests the advantages of trust without the potential confound of punishment. Because the investor will always accept the highest known, profitable offer before risking an unknown offer, the agent never punishes—the investor never rejects offers that are profitable ($r > 1/b$) in order to trust. Thus, in this simulation, we test whether trust is

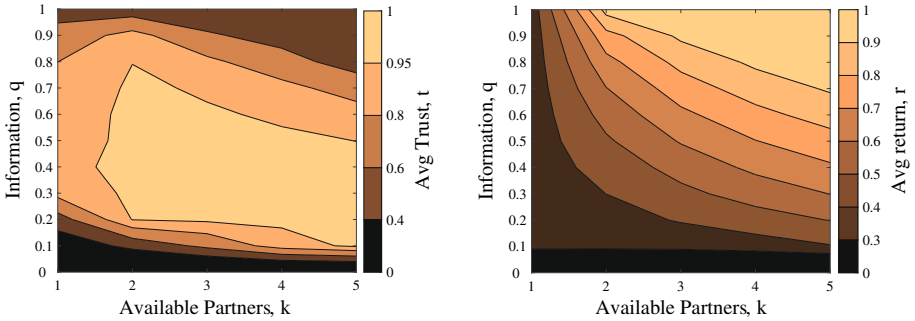


Fig. 1 Selection for trust in the investors (*left*) and return rate in the trustees (*right*), both as functions of information the investors hold about the trustees (q) and the number of partners available in each round (k). The depicted values are averages over 5 runs with populations of 500 averaged over the final 400 generations of each run. (*Left*) The average trust, t, in the investor population, where the investors accept the highest known return greater than $1/b$. (*Right*) The average return rate, r, in the trustee population

adaptive in a profit maximizing investor who never rejects profitable offers. In subsequent sections we will evaluate whether trust is adaptive when investors may reject profitable offers.

During each round, an investor is offered the opportunity to invest with one of k randomly selected trustees. Each investor plays $x = 500$ rounds of the game, and after x rounds, a new generation of investors and trustees are selected.[1] If an investor's trust (t) or a trustee's return rate (r) performs better compared to others, that agent and its attribute have a higher likelihood of appearing in the next generation.

After x rounds, the next generation of trustees and investors is simultaneously generated. In line with Manapat et al. [47], each agent is selected for the next generation using a variant of the pairwise comparison process [79]. During the creation of the next generation, each agent is randomly paired with another agent of the same type. The second agent adopts the strategy of the first with probability $\rho(1-\mu)$, where ρ is:

$$\rho = \frac{1}{1+e^{-\beta(\pi_a-\pi_b)}} \tag{2}$$

π_a and π_b represent the average payoffs of the first and second agent over x rounds. β affects the intensity of selection. We chose $\beta = 20$, as it is in line with Manapat et al. [47] and since it represents a relatively intense selection process, permitting new information and strategies to quickly permeate through the population. Finally, with probability $\mu = 0.01$, the second agent randomly selects a new strategy in the range [0, 1].

Initially, r and t are randomly instantiated in the range [0, 1], and throughout this work (until the final sensitivity analysis, Sect. 6), $b = 3$. When an attribute is added to the next generation, it is slightly mutated over a zero-mean, uniform distribution, ν, in the range $\nu \in [-0.005, 0.005]$. $g = 500$ generations are run, and the population's average trust and return rates are considered.

3.2 Results

Figure 1 shows a replication of the findings by Manapat et al. [47] with the novel decision rule described in Eq. 1. Figure 1(left) shows that the adaptiveness of trust depends on the market size (k), and the likelihood of possessing information (q). Generally, high levels of

[1] All models and code for figures are available in supporting materials.

trust evolve; investors are willing to invest with trustees despite ignorance of their individual return rate. However, a few points are worth highlighting. First, if the chance of possessing information is sufficiently low, then trusting is not advantageous. Further, trust requires more information as the market size decreases. When there is no partner choice ($k = 1$), trust begins to fail when $q < 0.3$. Finally, trust declines when the likelihood of information is high; however, this is not because trust is detrimental. Rather, trust is not needed so it is not selected for (see Sect. 3.3).

Figure 1(right) depicts the trustees' average return rate. When there is only one partner, the trustees return a rate barely greater than $1/b$. When partner choice is added ($k > 1$), trustee return rates become a function of both k and q. When either the number of partners (k) or the frequency of information (q) increases, so does the average return rate (r).

3.3 Discussion

Manapat et al. [47] showed that trust is adaptive as long as the chance of knowing at least one return rate is greater than $1/b$. Our results confirm this finding. In a one-shot game where an investor has no memory, or third-party reputational information about their partner, it can be adaptive to blindly trust another as long as the interaction operates in a market where information is occasionally available. Further, as the number of potential partners increases (i.e. the market grows), trust is adaptive even with lower levels of information.

It is interesting to note that when the amount of information is held static, trust seems to behave non-monotonically as the number of partners increases. When $k = 2$, trust is higher than when $k = 1$ for the same value of q. Then trust often declines as the number of partners increases further.

Trust at $k = 2$ is greater than when $k = 1$, because when $k = 1$, the average return rate of the trustee is barely above $1/b$. Thus, any fluctuations in return rates can reduce the value of trust because an unknown offer is worse than keeping the initial endowment. When $k \geq 2$, competition from partner choice keeps return rates high. Consequently, even when return rates fluctuate, trusting is typically advantageous.

In addition, we find that as information and partner choice increase there exists a threshold where trust appears to decline [cf. Fig. 1(left)]. This is not because trust is detrimental, but rather because trust is unnecessary. When the rate of information is high, the return rate of each potential partner is likely known; thus, the investor is rarely faced with the dilemma of trusting an unknown partner. Consequently, trust drifts neutrally; trusting and non-trusting investors perform similarly because they never need to trust [47].

Figure 1(right) depicts the underlying market competition between the trustees. A trustee can only receive money if it is selected for investment. Without partner choice ($k = 1$), the trustee offers the minimum value which is advantageous to the investor ($r > 1/b$). As both the number of partners and the frequency of information rise, the chance increases that an investor will learn more than one return rate. As such, the trustees must raise their return rates to compete for selection. When information is fully transparent ($q = 1$), trustees are forced to offer almost everything in order to outcompete other trustees.

This is in line with the work by Debove et al. [26], who showed in the ultimatum game that partner choice can lead to large returns when there is an imbalance in the number of investors and trustees. Again, however, whilst trustee partner choice may generate higher return rates, it does not explain the human propensity to reject all known profitable, but unfair offers. We address this in the next two experiments.

4 Costly Punishment Evolves Only with Partial Information

Whilst Manapat et al. [47] demonstrated that trust is adaptive under certain conditions, they could not explain the propensity of one-shot trust game players to reject all known profitable-but-unfair offers. In this section, we extend the Manapat et al. [47] model and demonstrate that such behaviour co-evolves with trust, partner choice, and partial information. We show that complete transparency of information does not lead to rejecting unfair offers. Rather, a demand for fairness evolves only when information is partially obfuscated.

In Experiment 1, an investor selected the largest known return rate, provided the rate was larger than $1/b$. As a consequence, investors were not permitted to reject a profitable offer. We can relate the above decision rule to the notion of minimal acceptable offer (MAO), found in the ultimatum game. As the name suggests, the MAO is the minimal offer an investor will consider.

In Experiment 1, investors' MAO was set to $1/b$. Here, we allow the MAO to evolve. By doing this, we enable selection for punishment which will sometimes manifest as costly. If an investor rejects all known offers greater than $1/b$, then the agent is rejecting profitable offers, which is a behaviour analogous to costly punishment witnessed in the one-shot trust and ultimatum game. The behaviour is in fact costly (at least local to the round) if either there are no unknown offers or the investor does not trust and accepts no other partner. We show that a MAO closer to fairness (1/2) is adaptive given partner choice and partial information.

4.1 Model

We add a new variable to each investor, demand $d \in [0 \ldots 1]$. Demand is the MAO that the investor will accept when a trustee's return rate is known. An investor is now characterized by both trust, t, and its minimum acceptable offer, d. If an investor's demand is 0.5, then it will only accept offers that are fair or better—the trustee must offer at least a 50% return. If the investor's demand is $1/b$, then it behaves exactly as before.

We can formalize the investor's decision rule. Based on the transparency of information, q, an investor knows the return rate for j out of k trustees. It invests with trustee i who has return rate r_i with probability:

$$p(r_i) = \begin{cases} 1 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x = r_i \geq d \\ 0 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x \neq r_i \\ \frac{t}{k-j} & i > j;\ \max\limits_{1 \leq x \leq j} r_x < d \end{cases} \tag{3}$$

There are only two changes from Eq. 1. First, a trustee with a known return rate is only chosen if its return is greater than or equal to d, rather than $1/b$. Second, based on its trust, t, an investor invests with an unknown trustee if none of the known return rates meet the MAO of the agent.

Since simulated evolution selects for the agents with the highest payouts, the decision rule creates pressure to find the revenue maximizing values of d, t, and r. This allows us to test whether rejecting profitable offers and trust can be adaptive in the limit case—when agents are attempting to maximize profit.

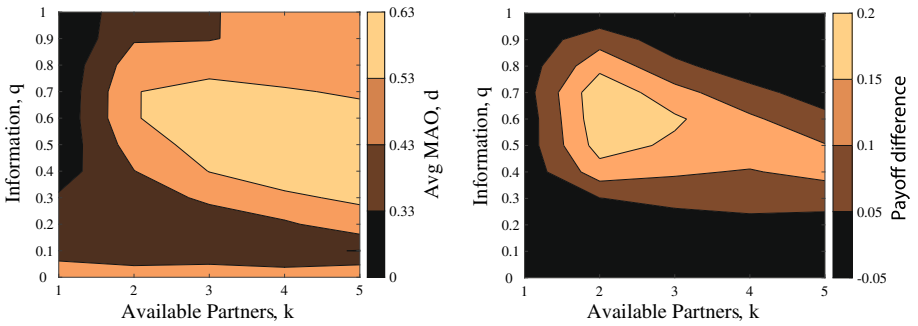


Fig. 2 Standards of fairness benefit investors and are selected for in conditions of partial information and partner choice. (*Left*) The investors' average minimum acceptable offer (d) over the final 400 generations, averaged over 5 runs. (*Right*) Investors' average fractional payoff difference when demand (d) evolves versus when it is held static at $1/b$, as in the first experiment

4.2 Results

Figure 2(left) depicts the population's average MAO (d) for differing values of q and k. Without partner choice ($k = 1$) and as long as information is moderately translucent ($q \geq 0.3$), it is not advantageous to demand a return rate greater than $1/b$. However, as soon as an investor can select between multiple partners, $k > 1$, then the average minimum acceptable offer increases. On average, investors are willing to reject profitable offers.

We then calculated the fractional payoff of each investor. The fractional payoff is calculated by taking an investor's average payoff across x rounds and dividing it by b. This represents the fraction of the maximum reward (in this case $b = 3$), which the investor received.

Figure 2(right) represents the fractional payoff difference between investors who evolved demand and those in the previous experiment where the MAO was held static at $1/b$. Positive values represent the contexts where an investor earns more by rejecting profitable offers. Thus, in Fig. 2(right), when the payoff difference is positive, then the ability to reject profitable offers (i.e. MAO $> 1/3$) leads to an increased payoff for the investor and consequently an evolutionary advantage. As the figure illustrates, it is only beneficial to reject profitable offers given partner choice ($k \geq 2$) and when information is not fully transparent ($q < 1$). Interestingly, these results, whilst reliable, derive from unstable dynamics—for an example run, see Fig. 3. This kind of highly cooperative, but unstable dynamic is often witnessed in the evolution of cooperation literature [22,36].

Finally, despite permitting the evolution of a MAO, trust still evolves. The graph is not shown due to space constraints, but the average levels of trust are similar to those depicted in Fig. 1(left).

4.3 Discussion

The tendency to reject all known profitable, but unfair offers in the trust or ultimatum game is often framed as a form of costly punishment. Here we show that such behaviour is profit maximizing in environments where information is partially obscured. In this model such behaviour is not necessarily costly punishment because individuals may in some rounds and conditions still accept another offer, which may possibly prove to be more profitable. As such, we show that costly punishment in these contexts may be accounted for by ongoing processes of negative pseudo-reciprocity [18], where players are able to punish partners by seeking other

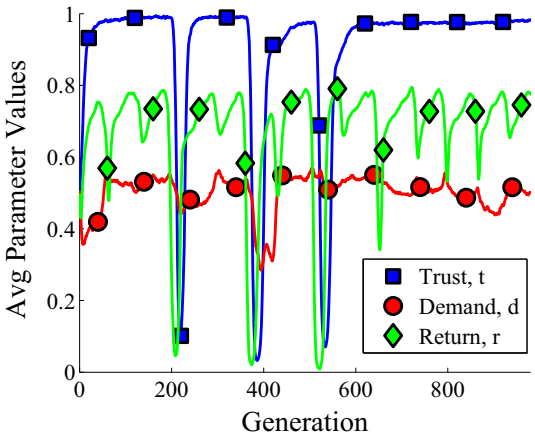


Fig. 3 Co-evolving average return rates (*diamonds*), trust (*squares*), and MAO (*circles*) over 1000 generations in a single exemplar run, where $k = 3$; $q = 0.5$. Trust and return rates are unstable, but reliably high on average; demand is similarly unstable but hovers near the fair value of 0.5. Defecting trustees occasionally benefit, before being culled

unknown pacts. Since humans typically operate in an environment with alternative (albeit unknown) options, rejecting all known profitable, but unfair options will sometimes be immediately beneficial, as well as providing long-term benefit by motivating higher rates of return.

Previous research has focused on analysing the evolutionary feasibility of either trust or costly punishment separately. We have shown that contexts exist where both blind trust and rejecting all known profitable, but unfair offers are simultaneously adaptive. In fact, possessing both a willingness to blindly trust and reject known profitable offers results in a better payout than either attribute in isolation. In small groups with partial information, there is selective pressure to both reject unfair offers and trust unknown offers.

Whilst rejecting profitable offers evolved throughout most of the parameter space (see Fig. 2(left)), such behaviour did not always increase investor payoff. For instance, when the number of partners and information prevalence are high [upper right corner of Fig. 2(right)], rejecting unfair offers is not advantageous. This is because with widespread knowledge of return rates (high information), competition between trustees pushes these rates to high levels. A willingness to reject offers of 0.5 is irrelevant if trustees are always offering returns above 0.9 [see Fig. 1(right)]. Thus, demand, d, has negligible effects and drifts neutrally.

Interestingly, there is only one context where rejecting profitable offers is advantageous—when there is both partially occluded information ($q < 1$), and partner choice ($k > 1$). What is unique about partial information? When information is not fully transparent, there is an opportunity for an investor to trust an unknown trustee. When information is transparent, trust is never a factor because, by definition, trust is only applicable when there is risk.

Our results indicate that rejecting profitable offers is adaptive because of its impact on the marketplace. The benefit derives from its impact on the selective landscape of the trustees. Rejecting all known offers is adaptive because it provides selective pressure on the trustees, forcing the trustees to increase their offers. Why is that? To further examine this phenomenon, we next explore demand rates when trust is removed from the investors.

5 Rejecting Profitable Offers is not Adaptive without Trust

Here we analyse the interdependence between trust and rejecting profitable offers. Above, we demonstrated that both blind trust and punishment behaviour that often manifests as costly

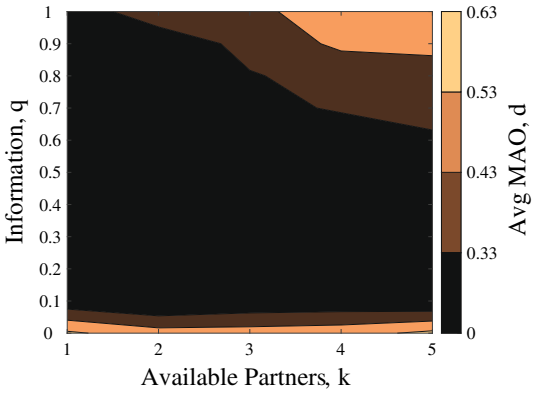


Fig. 4 A propensity to reject profitable offers cannot evolve without trust. Average investor MAO (d) is shown when trust is held static at 0. Results averaged over the final 400 generations and over 5 runs

(rejecting all known profitable-but-unfair offers) are adaptive with partner choice under partial information. Here we show that such behaviour is not adaptive without trust.

In the previous simulations, if no trustee offered a return rate above an investor's MAO, the investor had to decide whether to trust an unknown trustee, or keep its 1 unit of fitness. Here we force investor trust to zero. The probability of selecting a trustee i with return rate r_i now becomes:

$$p(r_i) = \begin{cases} 1 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x = r_i \geq d \\ 0 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x \neq r_i \\ 0 & i > j \end{cases} \tag{4}$$

When a trustee's return rate is unknown ($i > j$), it will never be selected. If no known trustee meets an investor's MAO, the investor will simply keep its one unit of fitness.

5.1 Results

Figure 4 shows the population's average MAO (d) when investors are untrusting ($t = 0$). Generally, rejecting profitable offers does not evolve; the MAO rarely raises above $1/b$. Only when both information and the number of investors are high is $d > 1/3$.

5.2 Discussion

Generally, rejecting profitable offers cannot evolve without trust. However, in the upper right corner of Fig. 4, higher minimum acceptable offers occur. This is, again, due to neutral drift in contexts where prevalent information leads to increased competition and higher rates of return. To validate that none of the evolved demand rates conferred advantages to the investor, we ran another simulation. As before, we held trust at zero ($t = 0$), but this time we also held the demand rate static at $d = 1/3$. In such a context, the investor will never leave a profitable offer on the table, but will also never trust an unknown offer. We subtracted the average payoff from investors who evolved d from those where $d = 1/3$. No evolved investor payoff outperformed investors where $d = 1/3$ (graph not shown, because all numbers were less than zero). Consequently, even when the average MAO drifts above $1/3$ in Fig. 4, it confers no benefit.

The finding that trust mediates adaptive costly punishment is congruent with the findings of Balliet and Van Lange [5], who show in a global meta-analysis that punishment only

promotes cooperation where levels of trust are high. However, there is a potential confound to their analysis—trust and cooperation are both correlated with wealth. It may be that trust, cooperation, and other public goods are a luxury more prevalent in societies well under the carrying capacity of their environment [74]. Nevertheless, the dynamics of our results hold across even the economically neutral contexts of abstract simulations, so must be considered a parsimonious explanation for this observed regularity.

In summary, rejecting profitable offers cannot evolve without trust. But, why is that? Should not high demand rates threaten the trustees? If trustees do not acquiesce to high demands, then no one will invest. However, we have shown that, in an evolutionary context, this threat is not sufficient to raise trustee returns. If simply raising demands was sufficient to increase returns, then high minimum acceptable offers would have evolved without trust.

What does trust confer which enables the rejection of unfair offers? High trust increases the likelihood that an investor will invest with a trustee despite ignorance of the trustee's return rate. If removing trust eliminates the advantages of rejecting profitable offers, then we know that it is not just the threat of the MAO which increases the return rates of the trustees. Instead, the combination of a high MAO mixed with the threat of investing with an unknown trustee increases the trustee return rates. By eliminating competition between known and unknown return rates, we have eliminated positive selection for d. Demanding a fair offer is only adaptive in the context of co-evolving levels of trust in environments of partial information.

6 Sensitivity Analysis

To validate that these results are not limited to a small parameter space, we ran a sensitivity analysis. We considered our results in the context of the number of partners (k), population size (N_i and N_t), the value of investment (b), selection intensity (β), and mutation rates (μ and ν). Generally, we found that our results hold for a wide range of environments. Blind trust and a willingness to reject unfair offers are robustly adaptive given partner choice and partial information.

6.1 Potential Partners: k

First, we tested whether our results hold as the number of potential partners increase (see "Appendix A"). When the number of partners (k) is extended to fifteen, our results continue to hold. When there is selective pressure for trust, it is advantageous to demand fair offers; and both behaviours require partial information and partner choice.

6.2 Population Size: N_i and N_t

Next, we tested whether our findings are sensitive to population size. In "Appendix B", we show that increasing the population of investors and trustees by tenfold ($N_i = N_t = 5,000$) does not alter the results.

6.3 Value of Investment: b

In "Appendix C", we tested whether our findings are sensitive to the value of investing, b. Above, when an investor selected a partner, the investment of 1 was multiplied by $b = 3$. We reran our simulations for $b = 2, 4$, and 5. Again, our main results held. There is always a state space where blind trust and demands for fairness are advantageous.

Interestingly, as b increases, both 1.) the state space and 2.) the relative benefit where both blind trust and demands for fairness are adaptive grow. These results offer some interesting predictions which seem to fall in line with the cultural and economic differences in the trust and ultimatum game. It is well known that industrially advanced, rich societies tend to demand levels of fairness much higher than other societies [40]. This model predicts that as the value of investment rises (such as in rich, industrial societies), it is more likely that society will operate in a state space where demands for fairness are adaptive. Of course, the trust game is a very simple metaphor for a marketplace, and other factors are assuredly affecting fairness norms; however, further research may prove enlightening.

6.4 Selection and Mutation: β, μ, and ν

It has been shown that selection intensity can alter the long-term benefits of costly punishment [65]. As such, we tested our results for differing values of β (see "Appendix D"). β represents the likelihood than an agent will copy a better performing strategy.

We found that blind trust and high MAOs co-evolve even when selection intensity is relatively weak. Whilst there is a threshold where our results fail, this is well below the empirical evidence which describes how often humans tend to adopt better performing strategies [20].

It has also been shown that cooperation can benefit from variation in the population [51]. As such, we ran a sensitivity analysis on our mutation rates, μ, and ν. As a reminder, μ is the likelihood that an individual will randomly change strategies during the selection phase. ν represents the accuracy with which a strategy is copied.

In the main text, $\mu = 0.01$ and $\nu = 0.005$. We tested $\mu = 0.002$ and $\mu = 0.05$ when $\nu = 0.005$. Further, we used $\nu = 0.025$ and $\nu = 0.001$, when $mu = 0.01$. Our results held in all instances (see "Appendix E"). Blind trust and rejecting unfair offers are adaptive given partner choice and partial information.

6.5 Robustness to Other Conditions Facilitating Cooperation

In separate work [17], we are exploring (i) using culture rather than evolution (imitation rather than reproduction), (ii) drawing both roles from the same population, such that each individual has an r, t, and d, and (iii) the impact of spatial structure. As might be expected from our results and from other existing literature (e.g. [72,77,78]), changes known to extend cooperation also extend cooperation in our model. This does not, however, necessarily mean that they increase trust. What we have shown is that conditions of high choice and high information favour informed cooperation, rendering trust moot. Where cooperation is fairly but not sufficiently reliable, high trust is selected; where there is insufficient choice to keep trust high, there may be moderate rates of trust and high demand. In some conditions cooperation is not supported at all. Ecosystem features that facilitate cooperation appear to shift all three phases of cooperation further into the low information, low choice parameter space.

7 General Discussion

We have proposed a novel explanation as to why humans trust unknown and thus potentially disadvantageous offers, yet reject profitable-but-unfair offers. In small markets with partial information, these behaviours together are both adaptive. This is because creating competition between trustees with known and unknown rates of return is advantageous for the investor.

The willingness to trust an unknown partner, occurring with the willingness to reject profitable-but-unfair known offers, generates just such a competition. Because investors are willing to risk unknown partnerships, they are able to evolve higher minimum acceptable offers. Once the competition between the unknown and known trustees is created, the trustees are forced to raise their return rates. Neither cooperation nor defection on the part of the trustees is stable, but overall cooperation is sufficiently frequent to generally provide high expectations for fairness. Even where these collapse, the dynamics of the market are such that the system rapidly recovers (Fig. 3). However, where investors do not trust unknown offers, raising minimum acceptable offers confers no benefit. Trust is a prerequisite for the evolution of rejecting profitable-but-unfair offers and, as suggested by Queller and Strassmann [61], requires a measure of ignorance.

The rejection of profitable-but-unfair offers when no alternative offer exists is a form of costly punishment and has historically been identified as such in one-shot trust or ultimatum games. Here we suggest a deflationary account for such behaviour. Rejecting all known profitable offers can be revenue maximizing in environments where the individual can select from other, unknown offers. An environment where information is also partially obscured is sufficient for the development of such behaviour. We demonstrate that behaviour which will sometimes manifest as costly punishment can be advantageous as long as there are also at least sometimes unknown offers which can be sought and trusted. There also needs to be enough information available for trust to evolve (see Model 1); however, if there is sufficiently low information, then rejecting profitable offers becomes adaptive (see Model 2). These conditions can be expanded via other mechanisms known to facilitate cooperation such as spatial structure, though some will also increase the range of conditions in which trust is not required for cooperation. In such conditions trusts' selection is therefore not supported, at least in this model.

It is well known that if investors can choose from several partners, then market competition can generate an expectation for fairness [9,34,70]. This idea has been discussed under the terms biological markets [6,55] and competitive altruism [35,68], and it has been empirically demonstrated that partner choice increases prosocial behaviour through competition [63,69, 80]. However, whilst partner choice explains why trustees would offer a higher return, such work does not describe why an individual would reject all available profitable-but-unfair offers.

Our results can be considered in the light of the recent discussion on how outside options affect fairness norms [4,26]. Debove et al. [26] show that the development of fairness norms depends on how an offer compares to other potential offers (i.e. an individual's outside options). In their work, demands for fairness developed as long as other partners were easily found. However, if finding another partner came at a high cost, then fairness norms failed to develop. In the present work, blind trust can be seen as increasing an investor's outside options. Rather than being detrimental, blind trust increases the number of potential investments, generating competition between unknown and known trustees. This, in turn, provides scope for rejecting all, known profitable, but unfair offers.

As previously mentioned, in our model the reason rejecting all profitable offers is revenue maximizing is because it does not always manifest as costly punishment. Individuals only pay a cost to reject the profitable offers when there are no unknown offers, when they do not trust unknown offers, or when an unknown offer is chosen but has a lower return rate than the rejected offer. In any one turn, investing may reduce their payoff with such a risk (thus paying a cost), but, on average, they turn a larger profit than by always accepting the highest known offer (see Model 2). Whilst the action of withholding cooperation in this context is therefore often costly punishment, it is always akin to negative pseudo-reciprocity, where

rejecting partnerships results in gains for the rejecting player. Costly punishment in one-shot trust and ultimatum games may best be understood in this same context.

To our knowledge, ours is the first work to consider the evolutionary feasibility of blind trust and negative pseudo-reciprocity simultaneously. Whilst McNamara and Leimar [52] showed that rejecting profitable offers can be advantageous when the player predicts increased payoffs with another partner, we have demonstrated how such a calculation interacts with the evolution of trust in a partially occluded environment. By including the natural assumption that information about others is partially occluded, we demonstrate how rejecting all known profitable-but-unfair offers is adaptive due to another human decision-making quandary, blind trust. Rejecting all known offers can be advantageous as long as the environment is transparent enough for trust to be adaptive, but not fully transparent. This is likely the environment most humans occupy.

Since costly punishment is known to confer long-term benefits in environments where reputation is built [41,71], a large amount of recent work on partner choice has considered the formation of multiple-round partnerships and the cost of leaving such partnerships [7,26]. It is important to note that the models presented here consider partner choice in a one-shot game. We do not consider partnerships which subsist over multiple rounds. Players have no memories and cannot form reputations. As such, this model parsimoniously describes experimental behaviour witnessed in the one-shot ultimatum and trust game. Rejecting unfair offers and blind trusting in a one-shot trust game does not require complex social dynamics. When a market operates in partial information, blind trust and demands for fairness fall out for free.

A potential criticism is that our results are considered over an evolutionary time frame. Humans are not genetically, unconditionally trustworthy; we frequently adjust our strategies based on prior experience [10,64,74]. Individuals are calculatingly trusting [81]. We believe that the present results can be considered over an individual lifetime, demonstrating the conditional nature of trust. Evolutionary algorithms are particularly useful in uncovering advantageous strategies in populations where the frequency distribution of strategies affect the outcome of each action [2]. They are learning algorithms and can be metaphorically applied to learning both within and across lifetimes [54]. If the evolutionary metaphor is ripped away, the algorithm still searches for the best strategy at any given moment in time. If each "generation" is interpreted as individuals attempting to find one of the best actions given the current state of the market, then Fig. 3 demonstrates the conditional nature of trust. Generally trust is advantageous; however, if trustees attempt to exploit investor trust, trust quickly dissipates. Importantly, trust just as quickly reappears when trustees begin to offer acceptable return rates. This is congruent with the Fudenberg et al. [31] finding that individuals will quickly forgive harmful transactions when it is in their benefit to do so.

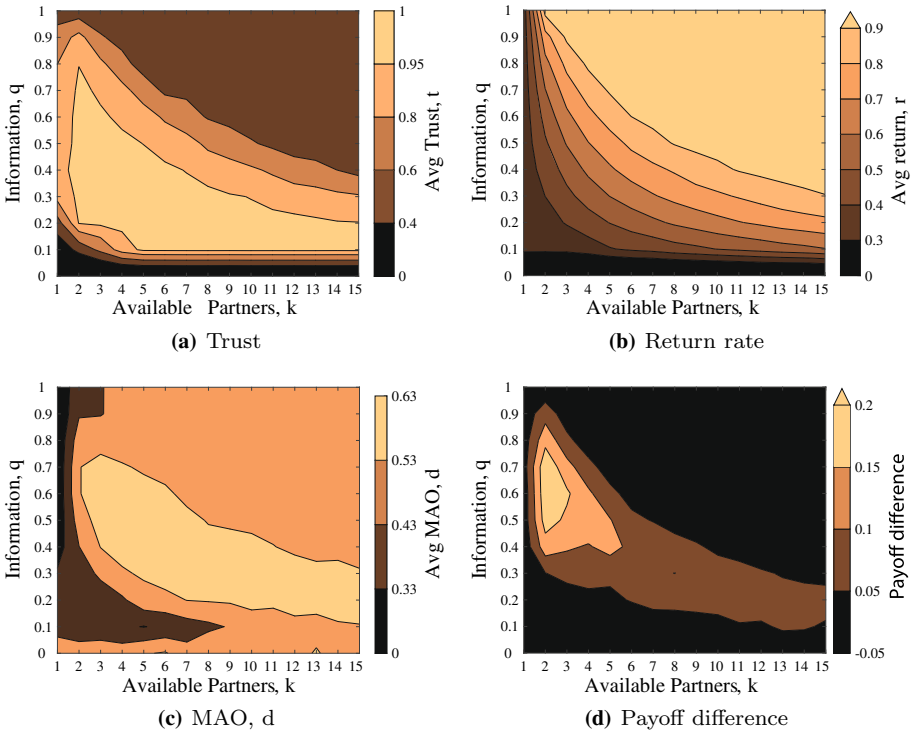
8 Conclusion

We have presented a parsimonious model accounting for both the rejection of profitable-but-unfair offers and the trusting acceptance of unknown offers. In environments of partial information where minimal acceptable offers and trust co-evolve, such behaviour is adaptive. In Experiment 1, we replicated the result that blind trust is adaptive provided there is sufficient information about one's partners. In Experiment 2, we demonstrated that rejecting profitable, unfair offers—a behaviour analogous to costly punishment in a one-shot ultimatum game—is advantageous presuming trust exists in environments of partner choice and information is

sufficient obfuscated. Finally, we showed that in such contexts, fairness expectations cannot evolve without trust. This provides a relatively simple explanation for both blind trust and the often costly rejection of unfair offers. Trust is adaptive given partner choice and partial information. Once trusting players frequent a population, rejecting profitable, but unfair offers is advantageous. Demanding fair offers cannot evolve without both ignorance and trust. Demanding a fair offer is only reliably advantageous if it includes the threat to accept unknown offers.

A Sensitivity Analysis: Number of Partners (k)

We ran a sensitivity analysis on the number of potential partners. In the main text, the number of partners ranged from $k = 1$ to 5. Here, we extend the analysis to include up to



(a) Trust **(b)** Return rate **(c)** MAO, d **(d)** Payoff difference

Fig. 5 Sensitivity analysis as partner choice increases to fifteen. Each value represents the population mean over the last 400 generations, averaged over 5 runs. b is held constant at three. **a** Average investor trust, t, when MAO (d) is held static at $1/b$. **b** Average trustee return rate, r, when MAO is held static. **c** The investors' average minimum acceptable offer, d, when it is permitted to evolve. **d** The difference between the investors' average fraction received when demand (d) evolves versus when it is held static at $1/b$

$k = 15$ partners. We ran two simulations. In the first, the investors' MAO (d) was allowed to evolve. In the second simulation, the MAO was held static at $1/b$. Figure 5 depicts the results. From $k = 1$ to 5, the graphs are identical to those in the main text.

Figure 5a illustrates the average trust when investor MAO is held static at $1/b$. Trust remains above 40% for all but the most obfuscated of environments ($q > 0.1$). When $q > 0.1$, levels of trust are high, but diminish given high information and a large number of partners. This is due to the fact that as the number of partners rise, the likelihood of knowing the return rates of multiple partners increases. When $k = 12 - 15$, this is true even in moderately obfuscated environments ($q \approx 0.5$). This causes trust to become superfluous. As mentioned in the main text, if enough information is known about a sufficient number of trustees, trusting unknown offers becomes unnecessary because a high offer is always known and available. Consequently, trust drifts near 50%.

Figure 5b depicts the average return rate of the trustee when MAO is held static at $1/b$. Compared to Fig. 1(right), the trend continues as expected. As the number of partners rise, return rates increase. This is due to competition between trustees. The highest return rates (white part of Fig. 5b) are directly correlated with eliminating selection pressure on trust (the dark brown portion of Fig. 5a). When a sufficient number of return rates are known, the average return rate rises towards one, and trust is no longer required.

Figure 5c illustrates the average MAO (d) when demand (d) is permitted to evolve. As in Fig. 2, demand is higher than $1/b$ whenever partner choice is greater than 1. Figure 5d depicts the fractional payoff difference between the simulation where d evolves, compared to when $d = 1/b$. As in Fig. 2(right), demanding a fair offer ($d > 1/b$) is only advantageous when information is partially occluded ($q < 1$), and when there are multiple partners ($k > 1$). However, the state space where demanding fairness generates an advantage diminishes as the number of partners rise.

As discussed in Sect. 5, this is because demanding a fair offer is only advantageous when there is selective pressure for trust. Since the utility of trust (t) relative to partner information (q) decreases as the number of partners (k) increase, rejecting unfair offers is advantageous in a smaller portion of the state space.

A.1 Discussion

The main finding of the paper continues to hold as the number of partners increase. When there is selective pressure for trust, then demanding fair offers ($d > 1/b$) is revenue maximizing. Blind trust and demanding fair returns co-evolve. As described in Sect. 5, pressure for the development of fair offers requires there is a benefit for blind trust.

One potential criticism concerns the decreasing state space where trust and demanding fairness co-evolve. If humans operate in large groups, then it becomes increasingly unlikely that society operates in the state space where both blind trust and high MAO co-evolve. Admittedly, this is a simple model and will not precisely project onto the complexities of the real world. However, we argue that humans have and do operate in a state space where trust and fairness norms are advantageous.

In order for trust to become superfluous and thus not co-evolve with fairness norms, market prices must be transparent enough that seeking an alternative price is rarely advantageous. This essentially amounts to a Bertrand competition [12], where perfect competition drives the prices of goods to the marginal cost. We suggest that this was not the case for most of human history. As a result, it would have been beneficial for individuals to develop both blind trust and demands for fairness.

The world and the market are, however, becoming increasingly transparent. This might have fascinating effects on the generation of trust and fairness norms in future societies. Whilst there is evidence that increased information transparency has not yet homogenized online price formation, nor reduced prices to marginal cost [25], the trend in both directions is apparent. Preliminary work as begun to consider how trust and fairness norms will change as societal transparency increases (see Bryson and Rauwolf [17]), but this remains an open question.

B Sensitivity Analysis: Number of Players (N_i and N_t)

Here the number of investors (N_i) and trustees (N_t) are increased to 5000, compared to the main text where $N_i = N_t = 500$. Our goal was to ensure that our results hold in a setting with significantly more players. We ran two simulations. In the first, investor MAO (d) was allowed to evolve. In the second, investor MAO was held static at $1/3$.

Figure 6 depicts the results. Figure 6a illustrates the average investor trust when MAO is held static at $1/b$. Figure 6b highlights the average trustee return rate when $d = 1/b$. Figure 6c shows the average investor MAO (d) when it is permitted to evolve. Finally, Fig. 6d depicts the payoff difference when investors evolve their MAO versus when it is held static. The

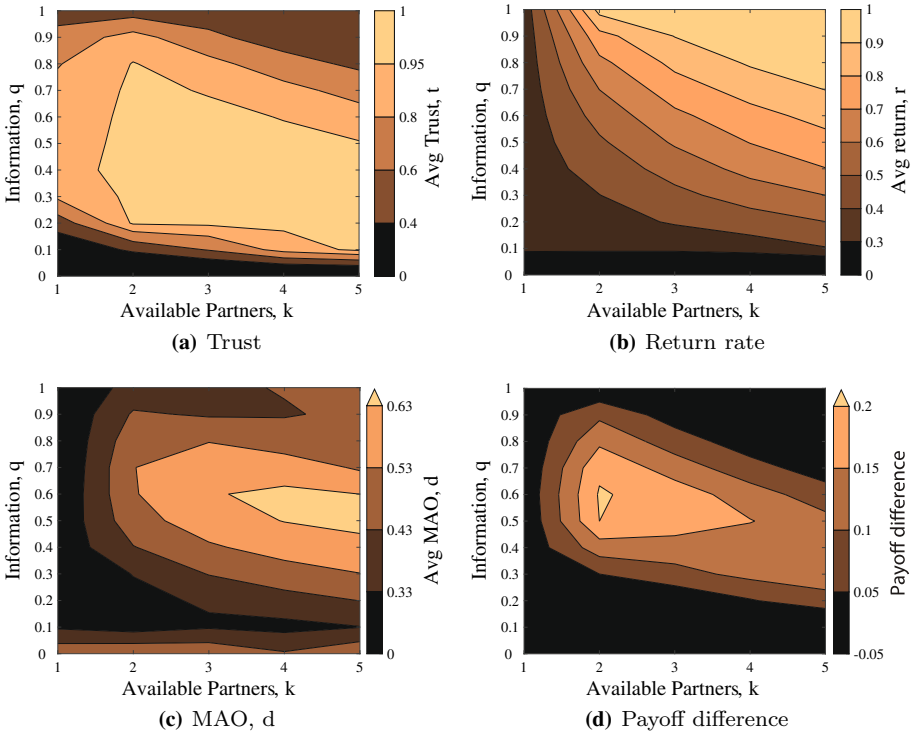


Fig. 6 Sensitivity analysis for 5000 investors and trustees. Each value represents the population mean over the last 400 generations, averaged over 5 runs. b is held constant at three. **a** Average investor trust (t) when MAO (d) does not evolve. **b** Average trustee return rate (r) when MAO does not evolve. **c** The investors' average minimum acceptable offer (d), when MAO evolves. **d** The difference between the investors' average fraction received when demand (d) evolves versus when it is held static at $1/b$

difference represents the fraction of the maximum reward ($b = 3$) that those with an evolved demand received over those where $d = 1/b$.

B.1 Discussion

Generally, there are few to no differences compared to the results of the main text. Trust (t), return rates (r), demand (d), and the payoff difference between the two models are analogous. The only time it is advantageous to demand fair offers is when blind trust co-evolves in an environment of partial information and partner choice.

C Sensitivity Analysis: b

In this section, we consider whether our results are sensitive to changes in b. b represents the value of investing; it is the scalar which is multiplied to an investment when an investor chooses to invest with a trustee. In the main text $b = 3$, so when an investor trusts a trustee, the investment of one is multiplied by three. Here we consider b =2, 4, and 5.

Two models were run for each value of b. First, we ran a model where MAO (d) was permitted to evolve. Second, we held MAO static at $1/b$. It is important note that altering b has a large effect on the notions of profit and fairness. An investor turns a profit as long as the trustee returns at least $1/b$ of the reward. When $b = 2$, 3, 4, and 5, this means an investor should reject all offers less than 1/2, 1/3, 1/4, and 1/5, respectively. When $b = 2$, the minimum profitable return is also a fair return, 1/2.

C.1 Results

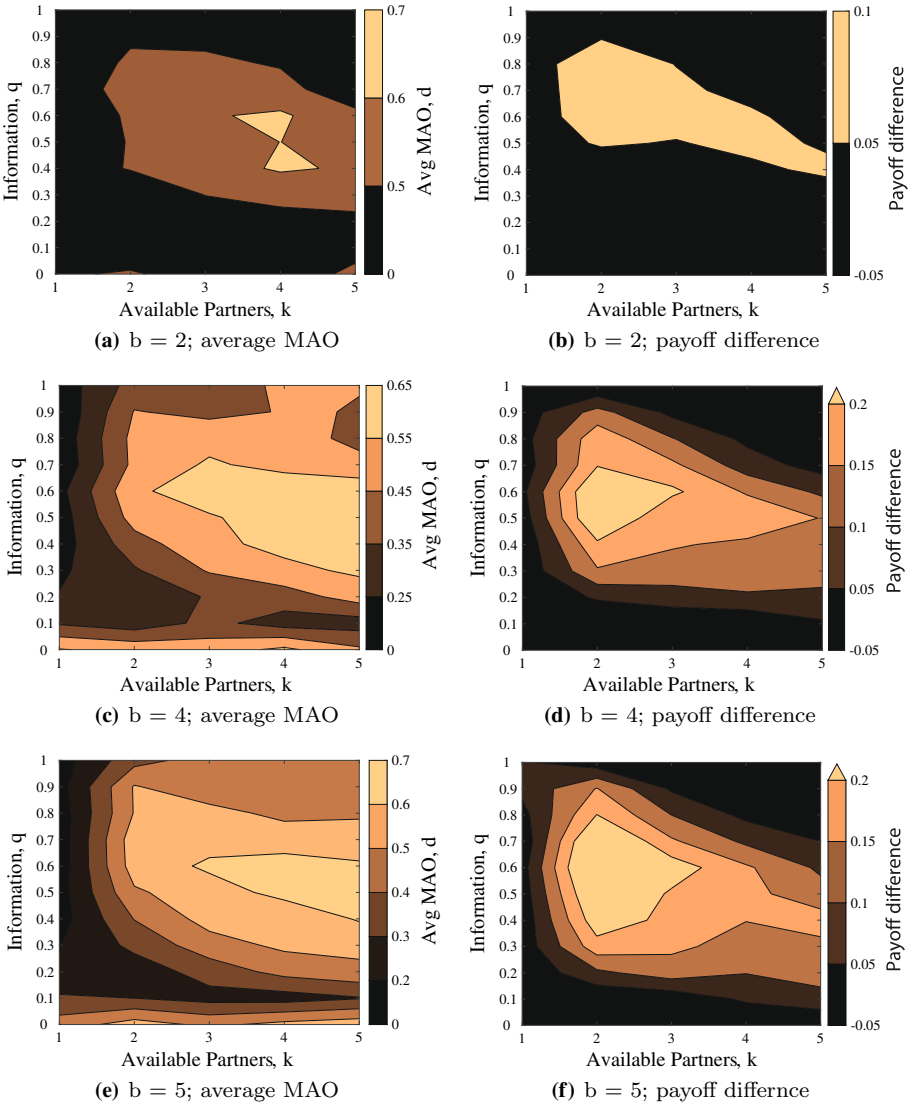
Figure 7 illustrates the results. The left graphs represent the average MAO (d) which evolved for various values of b. The right graphs represent the difference in the fraction of maximum reward (b) earned by an investor when MAO evolved, compared to when MAO was held static at $1/b$.

As mentioned above, the baseline for profitability shifts as b changes. The black shaded area on the left graphs represents demands which are worse than not investing. They change for different values of b. However, in all cases where trust is adaptive $q > 1/b$, rejecting profitable demands requires partner choice ($k > 1$).

The right graphs of Fig. 7 illustrate the payoff difference between those who evolved a MAO (d) and those where $d = 1/b$. The payoff difference represents the absolute payoff difference between the two groups divided by b. The payoff difference is calculated this way, so that we can compare the payoffs between different values of b. Obviously, when $b = 5$, agents are receiving more absolute money compared to when $b = 2$. By calculating the fraction of the maximum payoff (b) which is received, we can compare payoffs across differing values of b.

A few observations are immediately apparent. First, the results of the main text hold for all the values of b which we considered. The right graphs illustrate there is a state space where evolving demands above $1/b$ can be advantageous. This is true for all values of b in a space with partial information and partner choice. Trust and fairness norms co-evolve to produce a profit compared to other strategies.

The next observation is that the state space where blind trust and high MAO are adaptive grows as b increases. When $b = 2$, significantly less of the state space provides an advantage for blind trust and high MAOs. Relatedly, as b increases, the relative payoff grows between



(a) b = 2; average MAO **(b)** b = 2; payoff difference

(c) b = 4; average MAO **(d)** b = 4; payoff difference

(e) b = 5; average MAO **(f)** b = 5; payoff differnce

Fig. 7 Sensitivity analysis for b =2, 4, and 5. Each value represents the population mean over the last 400 generations, averaged over 5 runs. N_i and N_t are held constant at 500. (*Left graphs*) The average investor MAO (d), when MAO is permitted to evolve. (*Right graphs*) The difference between the investors' average fraction of the maximum reward received when demand (d) evolves versus when it is held static at $1/b$. **a**, **b** $b = 2$, **c**, **d** $b = 4$, **e**, **f** $b = 5$

those evolving demands and those who refuse to reject profitable offers ($d = 1/b$). It becomes more beneficial to blindly trust and reject unfair offers as the power of investment increases.

Interestingly, this not because higher fractions of payoffs are received by those who evolve their MAO. Instead it is because evolving the same MAO under different values of b is increasingly more profitable, compared to those who do not demand fair offers. Figure 8 illustrates this. Each subgraph represents the difference between the fractional payoff of those

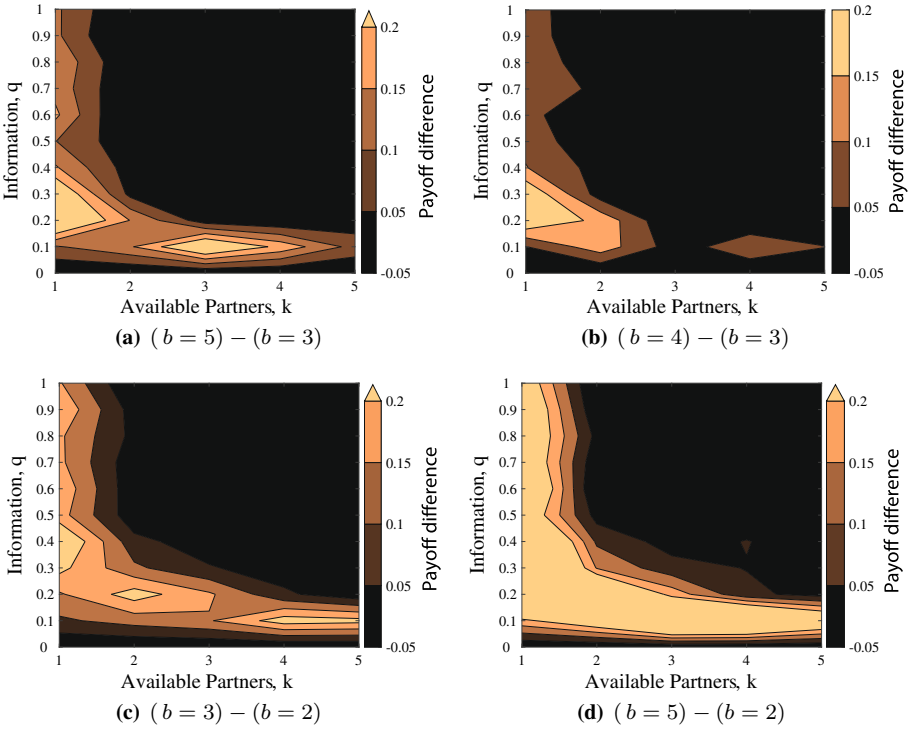


Fig. 8 Difference between the fractional payoffs of investors when MAO (d) evolves, for differing values of b. **a** Difference between investor payoff when $b = 5$ compared to $b = 3$. **b** Payoff difference between $b = 4$ and $b = 3$. **c** Payoff difference between $b = 3$ and $b = 2$. **d** Payoff difference between $b = 5$ and $b = 2$

who evolved their MAO for differing values of b. In the state space of partial information and partner choice where trust and MAO evolve, all four graphs are black. This indicates that the investors are receiving the same fraction of the payoff in all four cases.

Finally, it seems odd that in Fig. 7f, when $b = 5$ and there is no partner choice ($k = 1$), there appears to be a benefit to evolving demand (d) when information is high ($q > 0.9$). Previously, evolving demand was only advantageous given partial information ($q < 1$) and partner choice ($k > 2$).

We believe this is just an interesting artefact based on the difference in the variance of demand rates between the simulation which evolves MAO and the simulation where $d = 1/b$. It has been shown that variation in investor behaviour can raise trustee return rates [67], and when investor demands are permitted to evolve the variation in strategies is much higher than when all demands are held static at $1/b$. Importantly, however, this does not affect our main results. The average demand rate where $q > 0.9$ and $k = 1$ is less than $d = 1/b$ (see Fig. 7e). Thus, demands for fairness are not evolving, and slight increases in payoff are due to variation, not because agents are willing to reject profitable offers.

C.2 Discussion

In summary, the results of the main text hold for differing values of b. Partial information and partner choice lead to the co-evolution of blind trust and rejecting profitable offers ($d > 1/b$). Further, if b is permitted to evolve, then the average fraction of the reward received by the investor is analogous over different values of b (see Fig. 8). However, when comparing an investor who is willing to reject profitable offers ($d > 1/b$) to one who is not ($d = 1/b$), as the reward for investing increases (i.e. as b increases), 1.) the relative benefit, and 2.) the size of the state space for rejecting profitable offers increase.

These results offer some interesting predictions which seem to fall in line with the cultural and economic differences found in the trust and ultimatum game. It is well known that industrially advanced, rich societies tend to demand levels of fairness much higher than other societies [40]. This model offers a potential explanation for this. The relative advantages of trust and demands for fairness are far greater in rich, industrial societies where the reward for investment is much higher ($b >> 1$). Further, given that the state space increases with b, there is a higher likelihood that a wealthy, industrialized society will operate in a state space where blind trust and rejecting unfair offers confers an advantage. Of course, the trust game is a very simple metaphor for a marketplace, and other factors are assuredly affecting fairness norms; however, further research may prove enlightening.

D Sensitivity Analysis: β

Here we consider our results as β decreases. β represents the intensity of selection [79]. During the selection process, β manipulates the likelihood that one agent will adopt a better performing strategy (see Eq. 2). In the main text, $\beta = 20$. Consequently, an agent is highly likely to adopt a better performing strategy.

Here we consider $\beta = 10$, 5, 1, and 0.5. To put these values in perspective, imagine that Agent B is deciding whether to adopt Agent A's strategy. When $b = 3$, the average payoff for both agents rests in the range [0, 3]. Imagine that, on average, Agent A receives 0.3 more than Agent B (i.e. 10% more relative to the maximum payoff) each round. When $\beta = 20$, 10, 5, 1, and 0.1, then, using Eq. 2, the chance that Agent B will switch strategies is 0.9975, 0.9526, 0.8176, 0.5744, and 0.5374, respectively.

For each value of β, we ran two simulations. First, we allowed demand (d) to evolve. In the second, demand was held static at $d = 1/b$.

D.1 Results

Figure 9 depicts the results. Blind trust and high MAOs are adaptive for relatively weak levels of selection intensity. Though, as the likelihood of copying another's strategy decreases, the benefits of blind trust and demanding fairness diminish. This is to be expected, since the best strategy is not permeating through the society. However, even when $\beta = 1$, the advantage of demanding fair offers and blindly trusting is weaker, but it is still clearly visible (see Fig. 9c).

D.2 Discussion

When $\beta = 1$, there is less than a 60% chance that a strategy which performs 10% better will be adopted. This limitation is likely far below the threshold typically witnessed in

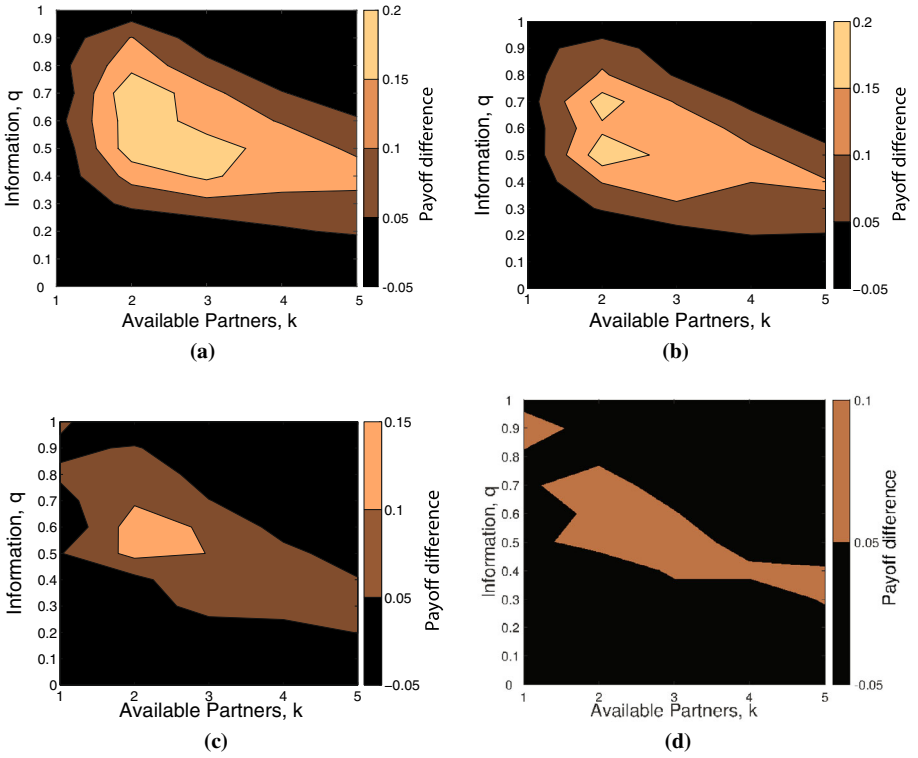


Fig. 9 Sensitivity analysis for β =10, 5, 1, and 0.5. Each graph represents the payoff difference between the investors' average fraction received when demand (d) evolves versus when it is held static at $1/b$. Each value represents the population mean over the last 400 generations, averaged over 5 runs. b is held constant at three. **a** $\beta = 10$, **b** $\beta = 5$, **c** $\beta = 1$, **d** $\beta = 0.5$

human societies. There is evidence that humans are much more likely to adopt advantageous strategies compared to this minimum threshold [20].

E Sensitivity Analysis: μ and ν

During the selection phase, there are two ways in which stochasticity is added to the model. First, with probability μ, an agent takes on a random strategy. Second, if an agent copies another's strategy, that strategy is uniformly mutated across the range $[-\nu, +\nu]$.

Here we run a brief sensitivity analysis on both μ and ν. In the main text $\mu = 0.01$ and $\nu = 0.005$. Here we test each parameter by multiplying the parameter by 5 and $1/5$. μ is tested at 0.05 and 0.002 whilst ν remains at 0.005. Then, ν is tested at 0.025 and 0.001 whilst $\mu = 0.01$. Two simulations were run for each test. In the first, MAO was permitted to evolve. In the second, $d = 1/b$.

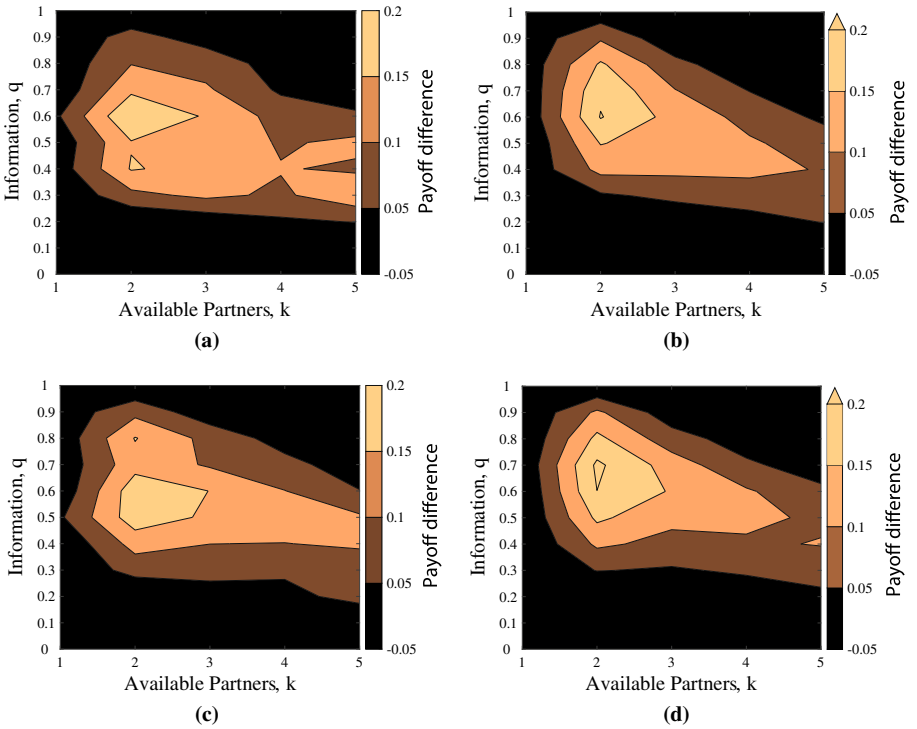


Fig. 10 Sensitivity analysis for μ and ν. Each graph represents the difference between the investors' average fraction received when demand (d) evolves versus when it is held static at $1/b$. Each value represents the population mean over the last 400 generations, averaged over 5 runs. b is held constant at three. **a** $\mu = 0.002$; $\nu = 0.005$, **b** $\mu = 0.05$; $\nu = 0.005$, **c** $\mu = 0.01$; $\nu = 0.001$, **d** $\mu = 0.01$; $\nu = 0.025$

E.1 Results

Figure 10 depicts the payoff difference when demand evolved, versus when it was held static. The left graphs illustrate the payoff difference when μ is changed. The right graphs depict the payoff difference when ν is altered.

E.2 Discussion

Generally, our results hold. Blind trust and the willingness to reject profitable offers co-evolve in environments of partner choice and partial information.

References

1. Al-Ubaydli O, Houser D, Nye J, Paganelli MP, Pan XS (2013) The causal effect of market priming on trust: an experimental investigation using randomized control. PLoS ONE 8(3):1–8. doi:10.1371/journal.pone.0055968
2. Alexander JM (2009) Evolutionary game theory. In: Zalta EN (ed) The stanford encyclopedia of philosophy, fall 2009 Edition
3. André J-B, Baumard N (2011a) The evolution of fairness in a biological market. Evolution 65(5):1447–1456

4. André J-B, Baumard N (2011b) Social opportunities and the evolution of fairness. J Theor Biol 289:128–135. http://www.sciencedirect.com/science/article/pii/S0022519311004012
5. Balliet D, Van Lange PAM (2013) Trust, punishment, and cooperation across 18 societies. Perspect Psychol Sci 8(4):363–379. doi:10.1177/1745691613488533
6. Barclay P (2016) Biological markets and the effects of partner choice on cooperation and friendship. Curr Opin Psychol 7:33–38. http://www.sciencedirect.com/science/article/pii/S2352250X15001906
7. Barclay P, Raihani N (2016) Partner choice versus punishment in human prisoner dilemmas. Evolution and Human Behavior 37(4):263–271. http://www.sciencedirect.com/science/article/pii/S1090513816000027
8. Barclay P, Stoller B (2014) Local competition sparks concerns for fairness in the ultimatum game. Biol Lett 10(5). http://rsbl.royalsocietypublishing.org/content/10/5/20140213
9. Baumard N, André J-B, Sperber D (2013) A mutualistic approach to morality: the evolution of fairness by partner choice. Behav Brain Sci 36(1):59–78
10. Bear A, Rand DG (2016) Intuition, deliberation, and the evolution of cooperation. Proc Natil Acad Sci 113(4):936–941. http://www.pnas.org/content/113/4/936.abstract
11. Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. Games Econ Behav 10(1):122–142. http://www.sciencedirect.com/science/article/pii/S0899825685710275
12. Bertrand J (1883) Théorie mathématique de la richesse sociale. Journal des Savants 48:499–508
13. Boero R, Bravo G, Castellani M, Squazzoni F (2009) Reputational cues in repeated trust games. J Socio-Econ 38(6):871–877. http://www.sciencedirect.com/science/article/pii/S1053535709000675
14. Bone JE, Raihani NJ (2015) Human punishment is motivated by both a desire for revenge and a desire for equality. Evol Hum Behav 36(4):323–330. http://www.sciencedirect.com/science/article/pii/S1090513815000203
15. Brañas-Garza P, Espín AM, Exadaktylos F, Herrmann B (2014) Fair and unfair punishers coexist in the ultimatum game. Sci Rep 4(6025). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4129421/
16. Bravo G, Tamburino L (2008) The evolution of trust in non-simultaneous exchange situations. Ration Soc 20(1):85–113. doi:10.1177/1043463107085441
17. Bryson JJ, Rauwolf P (2017) Trust, communication, and inequality. Extends and corrects a 2016 conference paper (in preparation)
18. Bshary R, Bergmüller R (2008) Distinguishing four fundamental approaches to the evolution of helping. J Evolution Biol 21(2):405–420. doi:10.1111/j.1420-9101.2007.01482.x
19. Burks SV, Carpenter JP, Verhoogen E (2003) Playing both roles in the trust game. J Econ Behav Org 51(2):195–216. http://www.sciencedirect.com/science/article/pii/S0167268102000938
20. Burton-Chellew MN, El Mouden C, West SA (2017) Social learning and the demise of costly cooperation in humans. Proc R Soc Lond B Biol Sci 284(1853). http://rspb.royalsocietypublishing.org/content/284/1853/20170067
21. Campenni M, Schino G (2014) Partner choice promotes cooperation: The two faces of testing with agent-based models. J Theor Biol 344:49–55. http://www.sciencedirect.com/science/article/pii/S0022519313005468
22. Cavaliere M, Sedwards S, Tarnita CE, Nowak MA, Csikász-Nagy A (2012) Prosperity is associated with instability in dynamical networks. J Theor Biol 299:126–138. http://www.sciencedirect.com/science/article/pii/S0022519311004619
23. Charness G, Dufwenberg M (2006) Promises and partnership. Econometrica 74(6):1579–1601. doi:10.1111/j.1468-0262.2006.00719.x
24. Chiang Y-S (2010) Self-interested partner selection can lead to the emergence of fairness. Evol Hum Behav 31(4):265–270
25. Clay K, Krishnan R, Wolff E, Fernandes D (2002) Retail strategies on the web: price and nonprice competition in the online book industry. J Ind Econ 50(3):351–367. doi:10.1111/1467-6451.00181
26. Debove S, André J-B, Baumard N (2015) Partner choice creates fairness in humans. Proc R Soc Lond B Biol Sci 282 (1808). doi:10.1098/rspb.2015.0392
27. Delgado MR, Frank RH, Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. Nat Neurosci 8(11):1611–1618
28. dos Santos M (2014) The evolution of anti-social rewarding and its countermeasures in public goods games. Proc R Soc Lond B Biol Sci 282:1798. http://rspb.royalsocietypublishing.org/content/282/1798/20141994
29. Engle-Warnick J, Slonim RL (2004) The evolution of strategies in a repeated trust game. J Econ Behav Org 55(4):553–573. http://www.sciencedirect.com/science/article/pii/S0167268104000721
30. Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Q J Econ 114(3):817–868. http://www.jstor.org/stable/2586885
31. Fudenberg D, David R, Dreber A (2012) Slow to anger and fast to forgive: cooperation in an uncertain world. Am Econ Rev 102(2):720–749. http://www.jstor.org/stable/23245432

32. Gintis H, Bowles S, Boyd R, Fehr E (2005) Moral sentiments and material interests: origins, evidence, and consequences. In: Gintis H, Bowles S, Boyd R, Fehr E (eds) Moral sentiments and material interests: the foundations of cooperation in economic life. MIT, Cambridge, pp 3–39
33. Güth W, Kliemt H (2000) Evolutionarily stable co-operative commitments. Theory Decis 49(3):197–222. doi:10.1023/A:1026570914311
34. Hammerstein P, Noë R (2016) Biological trade and markets. Philos Trans R Soc Lond B Biol Sci 371:1687. http://rstb.royalsocietypublishing.org/content/371/1687/20150101
35. Hardy CL, Van Vugt M (2006) Nice guys finish first: the competitive altruism hypothesis. Personal Soc Psychol Bull 32(10):1402–1413. http://psp.sagepub.com/content/32/10/1402.abstract
36. Hauert C, Holmes M, Doebeli M (2006) Evolutionary games and population dynamics: maintenance of cooperation in public goods games. Proc R Soc Lond B Biol Sci 273(1600):2565–2571. http://rspb.royalsocietypublishing.org/content/273/1600/2565
37. Hauser OP, Nowak MA, Rand DG (2014) Punishment does not promote cooperation under exploration dynamics when anti-social punishment is possible. J Theor Biol 360:163 – 171. http://www.sciencedirect.com/science/article/pii/S0022519314003920
38. Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, McElreath R (2001) In search of homo economicus: behavioral experiments in 15 small-scale societies. Am Econ Rev 91(2):73–78. http://www.jstor.org/stable/2677736
39. Henrich J, Ensminger J, McElreath R, Barr A, Barrett C, Bolyanatz A, Cardenas JC, Gurven M, Gwako E, Henrich N, Lesorogol C, Marlowe F, Tracer D, Ziker J (2010a) Markets, religion, community size, and the evolution of fairness and punishment. Science 327(5972):1480–1484. http://science.sciencemag.org/content/327/5972/1480
40. Henrich J, Heine SJ, Norenzayan A (2010b) The weirdest people in the world? Behav Brain Sci 33(2–3):6183
41. Hilbe C, Traulsen A (2012) Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. Sci Rep 2(458)
42. Hopfensitz A, Reuben E (2009) The importance of emotions for the effectiveness of social punishment*. Econ J 119(540):1534–1559. doi:10.1111/j.1468-0297.2009.02288.x
43. Huck S, Oechssler J (1999) The indirect evolutionary approach to explaining fair allocations. Games Econ Behav 28(1):13–24. http://www.sciencedirect.com/science/article/pii/S0899825698906911
44. Johnson ND, Mislin AA (2011) Trust games: a meta-analysis. J Econ Psychol 32(5):865–889. http://www.sciencedirect.com/science/article/pii/S0167487011000869
45. Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E (2005) Oxytocin increases trust in humans. Nature 435(7042):673–676
46. Krasnow MM, Delton AW, Cosmides L, Tooby J (2015) Group cooperation without group selection: modest punishment can recruit much cooperation. PLoS ONE 10(4):1–17. doi:10.1371/journal.pone.0124561
47. Manapat ML, Nowak MA, Rand DG (2013) Information, irrationality, and the evolution of trust. J Econ Behav Org 90(Supplement):S57–S75. http://www.sciencedirect.com/science/article/pii/S0167268112002296
48. Manapat ML, Rand DG (2012) Delayed and inconsistent information and the evolution of trust. Dyn Games Appl 2(4):401–410. doi:10.1007/s13235-012-0055-6
49. Marlowe FW, Berbesque JC, Barrett C, Bolyanatz A, Gurven M, Tracer D (2010) The 'spiteful' origins of human cooperation. Proc R Soc Lond B Biol Sci. http://rspb.royalsocietypublishing.org/content/early/2010/12/09/rspb.2010.2342
50. Masuda N, Nakamura M (2012) Coevolution of trustful buyers and cooperative sellers in the trust game. PLoS ONE 7(9):1–11. doi:10.1371/journal.pone.0044169
51. McNamara JM, Barta Z, Fromhage L, Houston AI (2008) The coevolution of choosiness and cooperation. Nature 451(7175):189–192. doi:10.1038/nature06455
52. McNamara JM, Leimar O (2010) Variation and the response to variation as a basis for successful cooperation. Philos Trans R Soc Lond B Biol Sci 365(1553):2627–2633. http://rstb.royalsocietypublishing.org/content/365/1553/2627
53. McNamara JM, Stephens PA, Dall SR, Houston AI (2009) Evolution of trust and trustworthiness: social awareness favours personality differences. Proc R Soc Lond B Biol Sci 276(1657):605–613. http://rspb.royalsocietypublishing.org/content/276/1657/605
54. McNamara JM, Weissing FJ (2010) Evolutionary game theory. In: Szkely T, Moore AJ, Komdeur J (eds) Social behaviour: genes, ecology and evolution. Cambridge University Press, UK, pp 88–106
55. Noë R, Hammerstein P (1994) Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. Behav Ecol Sociobiol 35(1):1–11. doi:10.1007/BF00167053

56. Perc M, Szolnoki A (2010) Coevolutionary games: a mini review. Biosystems 99(2):109–125. http://www.sciencedirect.com/science/article/pii/S0303264709001646
57. Pfattheicher S, Keller J (2014) Towards a biopsychological understanding of costly punishment: the role of basal cortisol. PLoS ONE 9(1):1–6. doi:10.1371/journal.pone.0085691
58. Pfattheicher S, Landhäußer A, Keller J (2014) Individual differences in antisocial punishment in public goods situations: the interplay of cortisol with testosterone and dominance. J Behav Decis Mak 27(4):340–348. doi:10.1002/bdm.1811
59. Pfattheicher S, Schindler S (2015) Understanding the dark side of costly punishment: the impact of individual differences in everyday sadism and existential threat. Eur J Pers 29(4):498–505. doi:10.1002/per.2003
60. Powers ST, Lehmann L (2013) The co-evolution of social institutions, demography, and large-scale human cooperation. Ecol Lett 16(11):1356–1364. doi:10.1111/ele.12178
61. Queller DC, Strassmann JE (2013) The veil of ignorance can favour biological cooperation. Biol Lett 9(6). http://rsbl.royalsocietypublishing.org/content/9/6/20130365
62. Raihani NJ, Bshary R (2011) The evolution of punishment in n-player public goods games: a volunteer's dilemma. Evolution 65(10):2725–2728. doi:10.1111/j.1558-5646.2011.01383.x
63. Raihani NJ, Smith S (2015) Competitive helping in online giving. Curr Biol 25(9):1183–1186. http://www.sciencedirect.com/science/article/pii/S0960982215002110
64. Rand DG, Greene JD, Nowak MA (2012) Spontaneous giving and calculated greed. Nature 489(7416):427–430. doi:10.1038/nature11467
65. Rand DG, Tarnita CE, Ohtsuki H, Nowak MA (2013) Evolution of fairness in the one-shot anonymous ultimatum game. Proc Natl Acad Sci 110(7):2581–2586. http://www.pnas.org/content/110/7/2581.abstract
66. Rankin DJ, dos Santos M, Wedekind C (2009) The evolutionary significance of costly punishment is still to be demonstrated. Proc Natl Acad Sci 106(50):E135. http://www.pnas.org/content/106/50/E135.short
67. Rauwolf P (2016) Understanding the ubiquity of self-deception: the evolutionary utility of incorrect information. Ch. 6. Heterogeneity in Costly Punishment Strategies is Adaptive, pp 100–105
68. Roberts G (1998) Competitive altruism: from reciprocity to the handicap principle. Proc R Soc Lond B Biol Sci 265(1394):427–431. http://rspb.royalsocietypublishing.org/content/265/1394/427
69. Roberts G (2015a) Human cooperation: the race to give. Curr Biol 25(10):R425–R427. http://www.sciencedirect.com/science/article/pii/S096098221500398X
70. Roberts G (2015b) Partner choice drives the evolution of cooperation via indirect reciprocity. PLoS ONE 10(6):1–11. doi:10.1371/journal.pone.0129442
71. Santos Md, Rankin DJ, Wedekind C (2010) The evolution of punishment through reputation. Proc R Soc Lond B Biol Sci 278(1704):371–377. http://rspb.royalsocietypublishing.org/content/278/1704/371
72. Sinatra R, Iranzo J, Gmez-Gardees J, Flora LM, Latora V, Moreno Y, (2009) The ultimatum game in complex networks. J Stat Mech Theory Exp 2009(09):P09012. http://stacks.iop.org/1742-5468/2009/i=09/a=P09012
73. Sylwester K, Herrmann B, Bryson JJ (2013) Homo homini lupus? Explaining antisocial punishment. J Neurosci Psychol Econ 6(3):167–188. doi:10.1037/npe0000009
74. Sylwester K, Mitchell J, Bryson JJ (2015) Punishment as aggression: uses and consequences of costly punishment across populations (in preparation)
75. Sylwester K, Roberts G (2010) Cooperators benefit through reputation-based partner choice in economic games. Biol Lett. http://rsbl.royalsocietypublishing.org/content/early/2010/04/13/rsbl.2010.0209
76. Sylwester K, Roberts G (2013) Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. Evol Hum Behav 34(3):201–206. http://www.sciencedirect.com/science/article/pii/S1090513812001250
77. Szolnoki A, Perc Mcv, Szabó G (2012) Defense mechanisms of empathetic players in the spatial ultimatum game. Phys Rev Lett 109:078701. https://link.aps.org/doi/10.1103/PhysRevLett.109.078701
78. Tarnita CE (2015) Fairness and trust in structured populations. Games 6(3):214–230. http://www.mdpi.com/2073-4336/6/3/214
79. Traulsen A, Pacheco JM, Nowak MA (2007) Pairwise comparison and selection temperature in evolutionary game dynamics. J Theor Biol 246(3):522–529. http://www.sciencedirect.com/science/article/pii/S0022519307000069
80. Van Vugt M, Iredale W (2013) Men behaving nicely: public goods as peacock tails. Br J Psychol 104(1):3–13. doi:10.1111/j.2044-8295.2011.02093.x
81. Williamson OE (1993) Calculativeness, trust, and economic organization. J Law Econ 36(1):453–486. http://www.jstor.org/stable/725485
82. Yamagishi T, Horita Y, Mifune N, Hashimoto H, Li Y, Shinada M, Miura A, Inukai K, Takagishi H, Simunovic D (2012) Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. Proc Natl Acad Sci 109(50):20364–20368. http://www.pnas.org/content/109/50/20364.abstract